A Human-Assisted Approach for a Mobile Robot to Learn 3D Object Models using Active Vision

Matthijs Zwinderman, Paul E. Rybski, Gert Kootstra

Abstract— **In this paper we present an algorithm that allows a human to naturally and easily teach a mobile robot how to recognize objects in its environment. The human selects the object by pointing at it using a laser pointer. The robot recognizes the laser reflections with its cameras and uses this data to generate an initial 2D segmentation of the object. The 3D position of SURF feature points are extracted from the designated area using stereo vision. As the robot moves around the object, new views of the object are obtained from which feature points are extracted. These features are filtered using active vision. The complete object representation consists of feature points registered with 3D pose data. We describe the method and show that it works well by performing experiments on real world data collected with our robot. We use an extensive dataset of 21 objects, differing in size, shape and texture.**

I. Introduction

Objects are the foundation of how humans interpret and reason about the world [1]. In order for robots to operate successfully alongside humans in real-world environments, robots too must be able to work with the objects that can be found within such environments. Object representations are often used in robotic systems as a basis of the world description, for instance for human robot interaction [2] or for acquiring the semantic structure of an environment [3]. However, where humans are experts at distinguishing individual objects regardless of pose or partial occlusion [1], object segmentation remains a challenge in current computer-vision and robotic systems. We wish to exploit this “expert knowledge” of object segmentation possessed by humans in order to create an effective human/robot interface which can transfer this knowledge to a representation that can be understood and exploited by a mobile robot.

Learning new objects that are appropriate for the tasks at hands is something we think that any robot that interacts with humans should be able to do. Teaching new objects should be possible on-the-fly with a fast and unambiguous interface that uses gestures or input methods that are natural and intuitive for the user.

In this paper, we propose a human-robot collaborative algorithm with which a human can refer to and teach novel objects using a simple protocol and an off-the-shelf laser pointing device. First we make sure that the robot and human have joint visual attention [2] to an object. After the human teacher designates the object, the robot actively explores it. The robot builds an object representation consisting of multiple views by integrating 3D data from stereo vision with interest points belonging to the object, this data is pruned using active vision to optimize recognition. The specific contributions of this work are the methods of using human knowledge to perform object segmentation in a natural fashion, methods to allow the robot to learn full 3D object models in a real-world environment with high levels of clutter. The system is shown to work well on an extensive test set of objects in an office environment.

II. Related Work

Referring to objects can be done in several ways in human-robot communication: by speech [4], by using an intermediary representation on a screen [5], [6], by deictic gestures [2] or by directly presenting the object to the robot [3], [7]. Combinations are also possible, such as the BIRON dialog management system [8], [9] where gesture information is used to resolve ambiguity of speech commands. When using speech to refer to objects, the robot needs to have prior knowledge about the object and the scene [5]. This is a very restrictive assumption, as the robot would need a large amount of prior knowledge about the world. We focus on learning novel objects in unknown environments.

The range of human input is often limited or restricted to make recognition computationally tractable. For speech, one can make a limited set of verbal commands [10] and to ease gesture recognition, limitations can be made on the positions and types of gestures, as well as other factors in the environment (such as the clothing of the human [9]). A large limitation on current gesture recognition is that the human needs to be in close proximity to the referred object [9], [8]. Natural gesturing is also limited in that it is not possible for the human to see whether the robot’s attention is guided to the correct position. Research by [11] shows that when gesture recognition systems misinterpret observed gestures, there is no clear way to disambiguate or correct the error. Specialized devices can be used to ease gesture recognition, such as the Xwand [12]. However, the Xwand requires that a 3D model of the room is created manually ahead of time and the pointing device itself is a specialized and complicated device. We propose to use a laser pointer, inspired by [5]. The laser pointer is a readily available device and our methods do not require models of the environment. The added benefit of a laser pointer is that it provides clear visual feedback to the human, making sure that the human and robot share visual attention on the same point in space.

Matthijs Zwinderman is with the Technical University of Eindhoven, The Netherlands mzwinderman@gmail.com

Paul E. Rybski is with the The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA prybski@cs.cmu.edu

Gert Kootstra is with the Royal Institute of Technology (KTH) in Stockholm, Sweden kootstra@kth.se

Fully automatic object segmentation is also possible, using active exploration [13], [14]. There remain some constraints on the objects and the environment: [13] is unsuitable for large objects and for both [13], [14] the object needs to be placed in front of the robot. Furthermore, an automatically created segmentation may not necessarily correspond to how a human perceives object segmentation. Like Kootstra *et al.* [14] we feel that it is very important for robots to have the ability to actively explore any environment and dynamically update their models accordingly, and we have therefor adopted their methods.

Once an object is segmented, an appropriate object representation must be created to allow for recognition on later observations. Object representations can simply consist of a set of images of the object, or a collection of features extracted from the object [15]. In the first case, objects can be recognized with techniques like normalized cross correlation [16] or neural networks [17]. A disadvantage of these matching techniques is that these techniques are often not invariant to light conditions, viewpoint changes and occlusions. To overcome such issues, object models based on invariant features such as SIFT [18] and SURF [19] can be used. In [14] SIFT is used in combination with active vision, to create an object model which holds different views of an object. In our approach, we use SURF feature points as SURF is faster and more robust than SIFT [19], we also store the 3D position of these points as computed by a stereo-vision system. By using a human teacher we can loosen the restraint placed on the environment in other works on active vision [14], [13], while keeping the advantages, such as recognizing objects from multiple poses.

III. Methodology

The process by which our robot learns an object representation starts with a person aiming a laser pointer at an object, see also figure 1. The robot's stereo camera identifies the reflection of the laser point projected on the object and tracks its position in 3D as the point moves around the object's surface (Step 1). After the human turns off the laser pointer, the robot uses the recorded positions of the laser pointer and the 3D information from the stereo camera of the object to segment the object from the background (Step 2). The robot then moves around the object and keeps its stereo camera pointed at the object in order to view it from all sides. Interest points and their position in 3D space are extracted from all viewpoints, to create a complete object representation (Step 3). By comparing different viewing angles, unstable interest points are discarded. Our object representation then is described as a set of multiple object-view pairs. These object representations are used to recognize the object in a later situation (Step 4).

A. Step 1 : Initial Segmentation of the Object from the Image

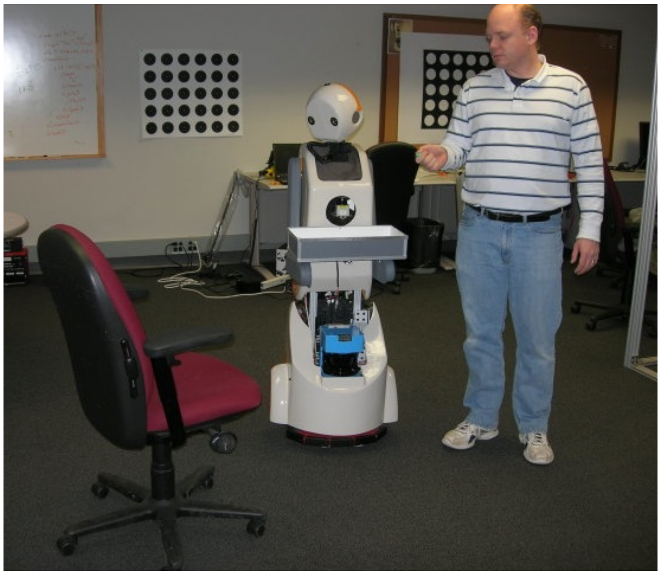


Fig. 1. A human teacher shows an object to the robot.

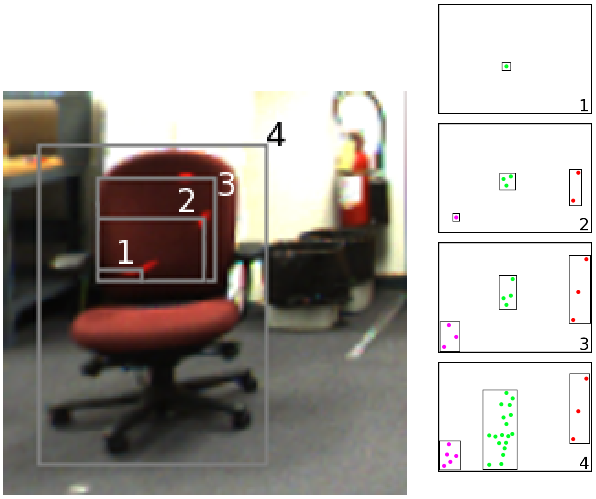


Fig. 2. After a new laser reflection is detected, its 3D-position is estimated using stereo vision and the detection is added to a cluster based on this position. The bounding box around the cluster with the most detections is overlaid on the original image on the left. The four boxes next to this show a side view of this clustering process, three clusters can be seen: the largest one representing the object. When the laser pointing stops, only the cluster with the most detections is kept, discarding the clusters with false detections. In this image four timesteps can be seen: from the first detection (1) to the final segmentation (4). The robot is positioned on the left.

The detection of the laser pointers is based on algorithms described by [5], [20] and consists of three steps: image differencing, color thresholding, and blob detection. Subtracting two subsequent images removes most of the image pixel data save for noise, reflections, and the projection of the laser pointer itself. The color thresholding then removes most of the noise and reflections. This thresholding process is only done on the red channel (since we're using a red laser pointer) and removes all pixels that are below an intensity value of 70. A morphological operation, dilation, is performed to reconnect laser reflections that were severed by the thresholding. Then a connected-components algorithm is run to find contiguous blobs. The largest blob that is smaller then 200 pixels (in a 320x240 image) is selected as the position of the laser point. The location in 3D space is calculated for each laser detection using a stereo vision system and the detections are clustered based on their 3D-position, taking co-variance into account. If the distance to existing clusters is too large a new cluster is created. Otherwise the new observation is added to an existing cluster. Clustering is necessary to filter out erroneous detections due to incorrect pointing, detection of reflections or incorrectly ascribing a detection to noise. The distance measure is the Mahalanobis distance in 3D space of the current detection to the centroid of existing clusters. The limiting distance is set to 0.05. The Mahalanobis distance is used due to the size differences of the objects. Figure 2 shows how the largest cluster of detections grows while new detections are added to it.

After 30 or more frames without a detection, the largest cluster is considered to be describing the object. During experimentation we have seen that successful segmentation is possible even if the path of the laser is not continuous. To extract an object mask from the laser data, a bounding box is created around the largest cluster of laser detections.

B. Step 2 : Initiation of the 3D Object Model

After the object has been segmented in the image space, a 3D bounding volume is created which represents an initial estimation of the volume of the object. Because the object occludes parts of itself from the robot, the *depth* of the bounding volume is unknown. We assume this to be equal to the *width* of the bounding volume, as seen from the robot. All dimensions of the bounding volume are grown by 20% of the originally estimated size to account for errors in the stereo-matching process and cases where the operator did not touch the boundaries of the object with the laser pointer. For our algorithm, it is more important to be conservative and create a larger initial bounding volume than to potentially exclude parts of the object, as interest points from the background are likely to be discarded by active vision (Step 3).

Our object representation consists of a collection of interest points generated by the SURF (Speeded Up Robust Features) algorithm [19]. SURF is an improvement upon the SIFT (Scale Invariant Feature Transform) [18] algorithm with enhancements in computational speed and robustness of the computed features. Such interest points are features in an image that can be easily matched and that are robust against rotation and lighting. The first set of interest points is generated from the first view of the object, which is described in step 3.

C. Step 3 : Iteratively Building the 3D Object Representation

After the initial object view is created and stored, the robot moves around the object while keeping its camera pointed at the object. The robot uses its odometry to update its world model, using its starting point as the origin. As new views of the object are captured, additional interest points are extracted. Using stereo vision the interest points are assigned a 3D location in the world model. Interest points that fall within the bounding volume set in step 2, are then added to the object representation as a new view. The use of odometry introduces an average error of 21.47 cm. in the estimation of the 3D position of interest points, but this error stays within the boundaries set around the bounding volume of the object [21]. Our algorithm is currently not responsible for directing the robot's motion and in our experiments the robot followed a pre-set path around the object. However, the world model is used to aim the robot's camera at a predetermined center of the object. It is fairly easy to let the robot adapt to the object's position and then do a predescribed circle.

An important challenge in this part of the algorithm is identifying and discarding unstable features; interest points that are only visible from one view or that are not robust against changes in viewpoint. Features are discarded by an algorithm described by [14] where interest points in a set N_n from view n are compared with interest points in a set N_{n-1} extracted from view n-1. We extended this method by also "looking ahead" to view n+1, allowing for greater changes in view point. Only those interest points that can be matched in either the prior or the subsequent view are kept, resulting in a set of filtered features F_n per view n.

$$\begin{aligned} F_n = & \\ & \{i \in N_n | \min_{x \in N_{n-1}} (||x - i||) < 0.6 \\ & \vee \min_{x \in N_{n+1}} (||x - i||) < 0.6\} \end{aligned} \tag{1}$$

Discarding unstable interest points has the advantage of lowering recognition time, as fewer interest points need to be matched. Recognition might even improve due to less noisy interest points in the object representation [14]. The angle between two subsequent viewpoints should not be too large, as then the mapping of interest points would fail due to the object obscuring itself partially and because SURF might not be robust against such a large change in viewpoint. We conducted a test was with 523 objects from the "Amsterdam Library of Image Objects" [22], showing that at a viewpoint change of 20° the SURF interest points can still be matched as reliably as SIFT at 10° (see [21]). We use the same algorithm as [14], but replace SIFT by SURF. In [14] a viewpoint change of 10° angle is used reliably, by using SURF this angle can be extended to up to 20°, reducing the computational burden.

D. Step 4 : Object Recognition

The recognition method is based on the activation model of Kootstra *et al.* [14]. The representation for a given object O is defined as a set of k views ω where k is a set of filtered interest points. An object can thus be defined as a set of object view pairs $\langle O, \omega_k \rangle$. Interest points detected in a new image are used to assign an activation level to each pair.

Let I be the collection of all SURF interest points extracted for a new image and D be the collection of all interest points in all object view pairs. For each interest point $o_i \in I$ the nearest neighbour interest point $d_{\langle O, \omega_k \rangle} \in D$ is calculated, using the Euclidean distance in interest point space. For this pair of interest points, an interest point activation value a_i is calculated as:

$$a_i = e^{-||o_i - d_{\langle O,\omega_k\rangle}||} \tag{2}$$

The total activation for the appropriate $\langle O, \omega_k\rangle$ is then updated according to the following formula:

$$A(\langle O, \omega_k\rangle | I, D) = \frac{\sum_i (a_i)}{\sqrt{|\langle O, \omega_k\rangle|}} \tag{3}$$

With $|\langle O, \omega_k\rangle|$ the number of interest points $\in \langle O, \omega_k\rangle$. This normalization step is done so that fewer matched observations are needed for objects that have relatively few interest points. This activation is summed over all views per object and the activation level results are ordered such that the higher total activation level reflects a higher certainty that the object is in the image. The highest scoring object is returned as the match. As suggested in [18] filtering of the interest points can also be done during recognition. However, this did not improve recognition scores in our case [21].

IV. Experimental Results

In order to test the accuracy and effectiveness of our algorithms, we performed two separate controlled experiments. The first experiment analyzed the performance and accuracy of object segmentation, the second experiment analyzed the ability of our algorithm to recognize objects after training.

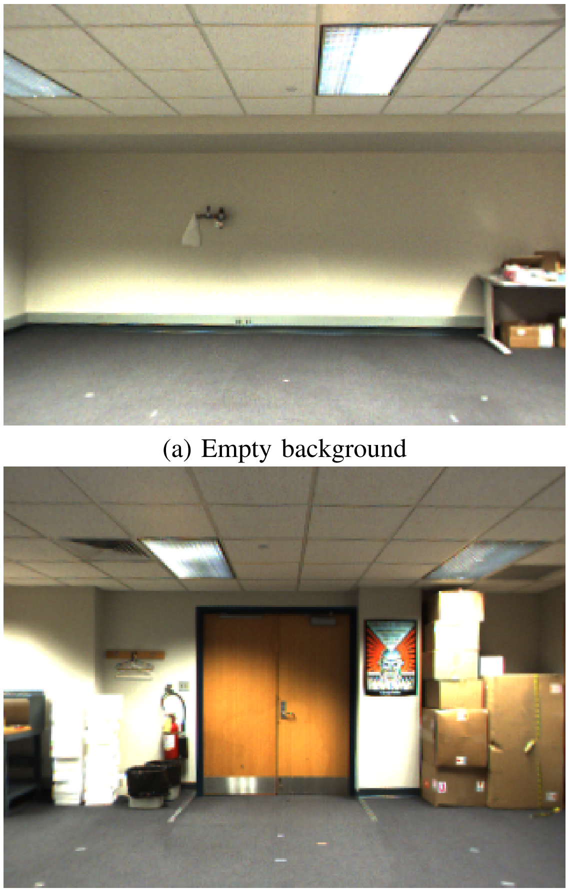
A. Robot Hardware and General Set-Up

The robot used in this work is a MobileRobots Pioneer Peoplebot, equipped with an on-board 1.6GHz Pentium-M single-board computer running Linux. The on-board computer was responsible for controlling the motion of the robot. The primary sensor used for this research was a Point Grey Research Bumblebee2 stereo camera. The Bumblebee2 is mounted on a Directed Perception pan-tilt unit model PTU-46-17.5. A Thinkpad X61 laptop with a 2.4GHz Intel Core 2 Duo processor was added to the robot to provide additional processing power for the stereo camera. The laser used by the person to point at objects in these experiments was a standard red laser pointer commonly used for slide presentations (class IIIa laser device with wavelengths 630-680nm).

The choice of background and objects has a huge influence on the performance of the stereo-vision calculations, as small and texturally sparse objects are too hard to distinguish using stereo vision. Care was taken to select differently sized objects with different textures to see how the methods are generalizable for the described datasets. The datasets were all created with one person pointing the laser pointer. Other research suggests no inter-subject differences when designating an object this way [5].

B. Object Segmentation Experiment

With this experiment the segmentation method was tested. The experimenter used the laser pointer to refer to a specific object in space, by moving the laser reflection over it. The object segmentation dataset consists of seven objects from an office environment (red chair, table, dust bin, cabinet, box, monitor and an arm chair see figure 4), with two different backgrounds: a simple and a complex one. The simple background consists of a plain wall, while the complex background has multiple other objects placed in it. All objects were placed in the robot's view in three different locations, one at time. On each location the object was recorded twice, changing the angle slightly for each run. The recordings lasted thirty seconds, during which the object was designated with the laser pointer.


(a) Empty background

(b) Complex background

Fig. 3. The two backgrounds used in the segmentation experiment

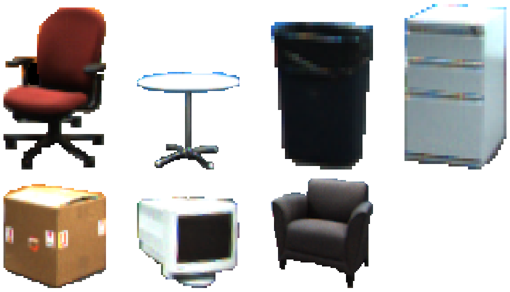


Fig. 4. Objects in the segmentation set.

The accuracy of the segmentation method was tested against a manually created ground truth. To measure the results of the algorithms, we used the True Positive Rate (TPR) and False Positive Rate (FPR), by comparing pixels

TABLE I

The TPR and FPR for the two backgrounds. The high TPR for the segmentation method shows that objects are segmented correctly using human guidance.

Background	TPR	FPR
Simple	$\mu = 0.943$ $\sigma = 0.138$	$\mu = 0.033$ $\sigma = 0.028$
Complex	$\mu = 0.854$ $\sigma = 0.294$	$\mu = 0.123$ $\sigma = 0.146$

in the ground truth image against the automatic segmentation. The results of the experiments were averaged over all objects, poses and positions and can be found for both backgrounds and segmentation methods in table I. The high TPR and low FPR, shows that the segmentation method is working very well for both backgrounds.

C. Object Learning Experiment

To test the creation of object representations and the recognition method, 21 office-environment objects were used, of which a selection is shown in figure 5 (for the complete set, see [21]), these objects differ greatly in size and texture complexity. The learning dataset was recorded with the robot, for each object the robot started at a fixed location, three meters from the object. Each trial started with the person painting the object with the laser for thirty seconds to show the robot what to learn. After segmenting the view of the object from the background, the robot would drive forward to a distance of 1 meter from the object and proceed to circle the object, stopping to collect images at eighteen predefined locations distributed evenly on this circle. A test set was made in the same manner, only here the distance of the robot to the object was varied slightly per location. For the scope of this paper, we were interested only in the mechanism for learning the object model and the robot base was driven manually around the object by the experimenter. The robot did automatically aim its camera at the predetermined center of the object at all times. Each of the eighteen viewpoints differs by up to 20 degrees with the previous one.

A representation was created for each object using the method described in III-C, termed "Active Filtering". To evaluate the recognition performance of this representation, two other types of representation were created. One without active vision filtering ("No-Active Filtering"), to evaluate the influence of this filtering technique. And one where only one viewpoint was used to create an object model (single view approach), as this type of object representation is often used in other robotic systems. As one viewpoint might be more representable of an object than another, each viewpoint was used as an object representation in the single view approach and the recognition performance was averaged over all viewpoints.

Each object in the testset was recorded from multiple viewpoints, and for each of these views the recognizer was run three times, for each of the types of object representations. The recognizer scored 1 for a correct recognition and 0 for incorrect. For the single-view approach, the scores were averaged over all views. Figure 6 show the average recognition performance per type of object representation. As can be seen, active filtering delivers the best recognition performance and the results were found to be border-line significant compared to no active filtering using a Student's t-test ($p = 0.05, \alpha = 0.05$). The main advantage of active filtering is that a large amount of useless data can be scrapped by filtering using active vision. The total amount of interest points for representations that were actively filtered is 82.87% that of the total amount of interest points in representations that were not filtered, resulting in faster recognition times.

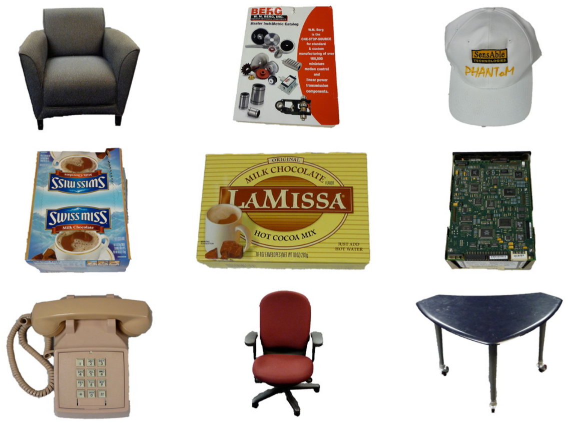


Fig. 5. A selection of the 21 objects in the object-learning and -validation set, the objects are all part of a normal office environment and are diverse in size and texture

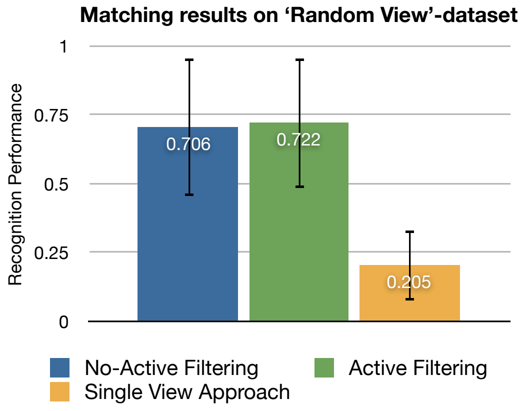
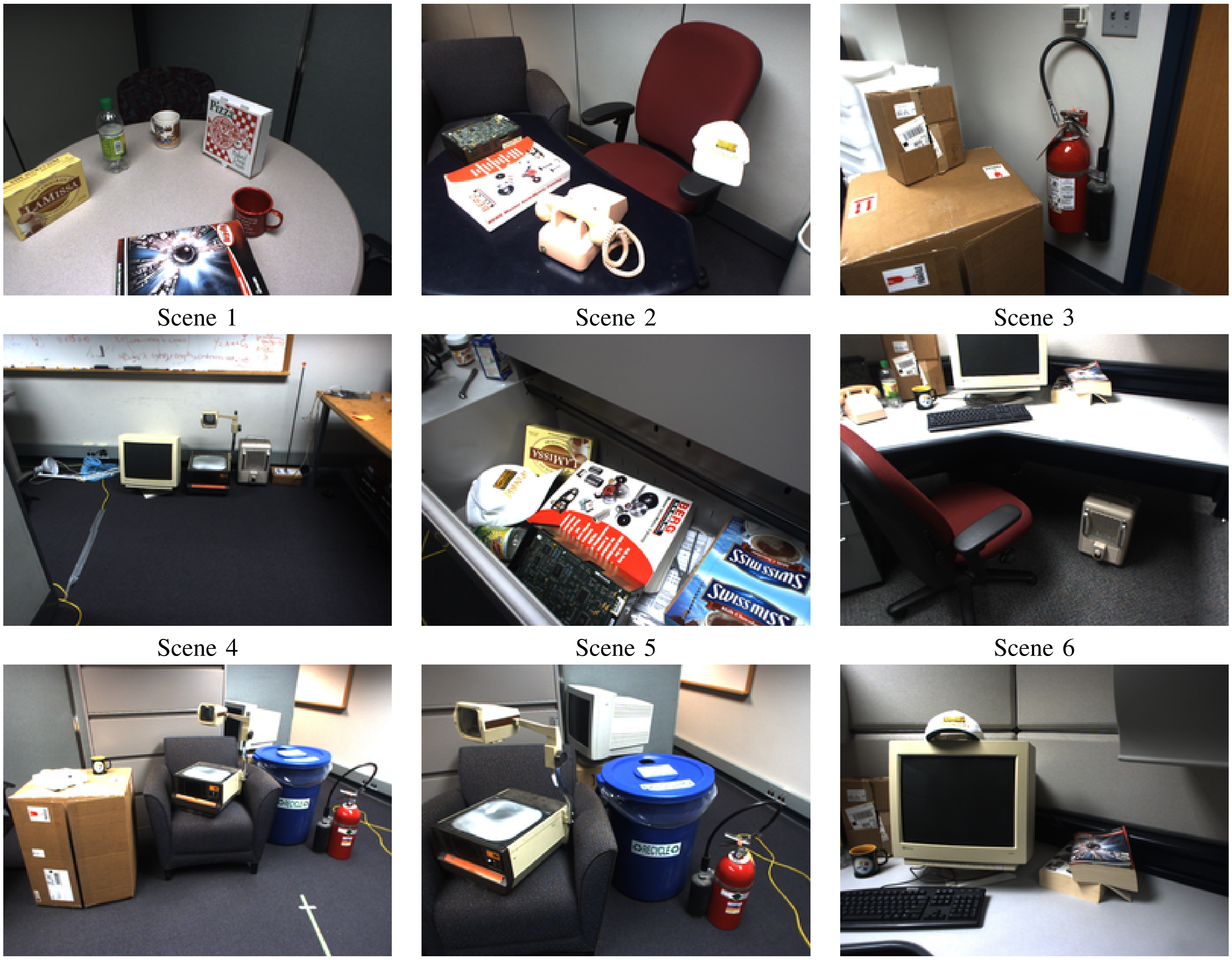


Fig. 6. The recognition results for each type of object representation

To test the learned object models in real-world situations, we generated seven different office scenes using the objects that were learned. The lighting differs per scene and objects can be partially occluded, making recognition a challenge. From each scene multiple observations were made called trials, each from a different angle. For each trial t the objects in the scene were transcribed in a set T_t. Examples of such scenes can be found in figure 7. We ran our recognizer on each t with object representations with and without



Scene 1

Scene 2

Scene 3

Scene 4

Scene 5

Scene 6

Scene 7

Scene 7 (different trial)

Scene 6 (different trial)

Fig. 7. Nine examples of different scenes and trials of objects in a natural setting. These objects come from the 'object-learning and validation-'-set, a subset of which is shown in figure 5

active filtering and on single view object representations. The recognizer returns a list L_t of matching objects ordered by activation. The number of correct objects per trial t is then $\#((O_1 \cdots O_n \in L_t) \cap T_t)$, with $n = \#T_t$.

Over all trials there are 92 objects. When object representations created without active filtering were used, the recognizer correctly recognized 53 objects, with filtering 52 objects and in the single view approach 28 objects. Again there is an obvious advantage for using a multiple views object model. The difference between the active and passive filtering is not significant. Overall, only a little over half of the objects are recognized, but these include objects that are only partially visible or that are observed from a completely different viewpoint than in which they where learnt. Several trials include oversaturation of the CCD-sensors due to artificial lighting. Examples of this can be seen in figure 7, scene 2 trial 1 and scene 6 trial 4. Here the phone object, as well as other parts of the image, is partially oversaturated. For these white pixels it is impossible to calculate meaningful keypoints. This is a limitation of the camera and especially of its gain selection algorithm. The problem is considerably worsened due to the bright uniform fluorescent lighting of the robot lab. A solution for this would be to perform the recognition on multiple images per location, using a different gain setting for each image.

V. Discussion

In this paper we presented a human-robot cooperative method for object learning, consisting of two parts: human segmentation and object learning with active vision. We have shown that by extending method of [5], a laser pointer can be used to successfully designate and segment objects in three-dimensional space. Our method performs very well, having a high positive rate and very low negative rate. By using a laser pointer the segmentation is accurate over a distance of three meters, unlike natural gesturing which has to be done in close proximity [9], [8]. Our method does not require an extensive world model, as is needed for speech [10] or other assisted gesturing methods [12]. The laser pointer also provides clear visual feedback to the user, which is not present in speech or natural gestures. We show that our segmentation method provides an accurate starting point for successful object learning.

We compared our active method with two typically used types of object representations (a single view and a non-filtered representation) in situations within which our robot should be able to operate. With the proposed methods it is possible for a robot to build robust object representations, and recognize an extensive set of 21 objects in 72.22% of the cases in an environment similar to the learning environment. Compared to our baseline single-view approach, the use of multiple views significantly increases recognition rate and compared to the passive approach the amount of data in our representations is considerably less, while the recognition rate does not decrease. Even in challenging complex scenes, more than half of the objects were recognized. It is difficult to compare our results with results from similar research due to differences in robot-platform, environment and dataset.

For our future work, we plan to improve our system in several ways. Currently we store interest points per view, making it possible to detect objects from multiple viewpoints and to perform pose recognition. As of yet, we do not use the configuration of the interest points in 3D space, while this can be used to gain more accurate pose recognition and increase the overall recognition rate [18], [23]. We also expect to drastically improve recognition rates by using active exploration [14], [24]. Our object representations already contain the necessary information for these extensions.

We have presented a complete system with which a human teacher can easily and reliably segment objects varying in size and structure from a complex background, to teach new objects to a mobile robot. We implemented this system on a mobile robot and have shown with experiments on real-word data that our system is indeed successful at segmenting, learning and recognizing objects.

Acknowledgments

We gratefully acknowledge the support of the National Science Foundation, through grant CRI 0709077, and Microsoft Research. We would also like to thank Jaldert Rombouts and Tijs Zwinkels for their help with the software architecture and low-level stereo camera drivers.

References

[1] J. Feldman, “What is a visual object?” *Trends in Cognitive Sciences*, vol. 7, no. 6, pp. 252–256, 2003.

[2] A. G. Brooks and C. Breazeal, “Working with robots and objects: revisiting deictic reference for achieving spatial common ground,” in *HRI ’06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. New York, NY, USA: ACM, 2006, pp. 297–304.

[3] S. Ekvall, P. Jensfelt, and D. Kragic, “Integrating active mobile robot object recognition and slam in natural environments,” in *Proceedings of the IEEE/RSJ International Conference on Robotics and Automation*, 2006, pp. 5792–5797.

[4] S. Wachsmuth, G. A. Fink, F. Kummert, and G. Sagerer, “Using speech in visual object recognition,” in *Mustererkennung 2000, 22. DAGM-Symposium Kiel, Informatik Aktuell*. Springer, 2000, pp. 428–435.

[5] C. C. Kemp, C. D. Anderson, H. Nguyen, A. J. Trevor, and Z. Xu, “A point-and-click interface for the real world: laser designation of objects for mobile manipulation,” in *HRI ’08: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. New York, NY, USA: ACM, 2008, pp. 241–248.

[6] C. Leroux, I. Laffont, N. Biard, S. Schmutz, J. F. Desert, G. Chalubert, and Y. Measson, “Robot grasping of unknown objects, description and validation of the function with quadriplegic people,” in *Rehabilitation Robotics, 2007. ICORR 2007. IEEE 10th International Conference on*, 2007, pp. 35–42.

[7] H. Wersing, S. Kirstein, M. Gotting, H. Brandl, M. Dunn, I. Mikhailova, C. Goerick, J. Steil, H. Ritter, and E. Korner, “Online learning of objects in a biologically motivated visual architecture,” *International Journal of Neural Systems*, vol. 17, no. 4, pp. 219–230, 2007.

[8] I. Toptsis, S. Li, B. Wrede, and G. A. Fink, “A multi-modal dialog system for a mobile robot,” in *Proc. Int. Conf. on Spoken Language Processing*, vol. 1, 2004, pp. 273–276.

[9] S. Ghidary, Y. Nakata, H. Saito, M. Hattori, and T. Takamori, “Multi-modal interaction of human and home robot in the context of room map generation,” *Autonomous Robots*, vol. 13, no. 2, pp. 169–184, 2002.

[10] S. Tomko and R. Rosenfeld, “Speech graffiti vs. natural language: Assessing the user experience,” in *in Proceedings of HLT / NAACL 2004*, 2004, pp. 73–76.

[11] J. Ziegler, K. Nickel, and R. Stiefelhagen, “Tracking of the articulated upper body on multi-view stereo image sequences,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 774–781, 2006.

[12] A. Wilson, S. Shafer, and S. Shafer, “XWand: UI for intelligent spaces,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM New York, NY, USA, 2003, pp. 545–552.

[13] P. Fitzpatrick, “From first contact to close encounters: A developmentally deep perceptual system for a humanoid robot,” Ph.D. dissertation, Massachusetts Institute of Technology, 2003.

[14] G. Kootstra, J. Ypma, and B. de Boer, “Active exploration and keypoint clustering for object recognition,” in *IEEE International Conference on Robotics and Automation, 2008. ICRA 2008*, 2008, pp. 1005–1010.

[15] S. Dutta Roy, S. Chaudhury, and S. Banerjee, “Active recognition through next view planning: a survey,” *Pattern Recognition*, vol. 37, no. 3, pp. 429–446, 2004.

[16] C. Rother, V. Kolmogorov, and A. Blake, “Grabcut : interactive foreground extraction using iterated graph cuts,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, August 2004.

[17] A. Gopalakrishnan, S. Greene, and A. Sekmen, “Vision-based mobile robot learning and navigation,” in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, 2005, pp. 48–53.

[18] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[19] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *Computer Vision – ECCV 2006*, 2006, pp. 404–417.

[20] D. Olsen Jr and T. Nielsen, “Laser pointer interaction,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM New York, NY, USA, 2001, pp. 17–22.

[21] M. J. Zwinderman, “Object Learning and Recognition using Human-Guided Object Segmentation and Active Vision,” Master’s thesis, University of Groningen, 2009.

[22] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, “The Amsterdam library of object images,” *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.

[23] E. Murphy-Chutorian and J. Triesch, “Shared features for scalable appearance-based object recognition,” in *Application of Computer Vision, 2005. WACV/MOTIONS ’05 Volume 1. Seventh IEEE Workshops on*, vol. 1, 2005, pp. 16–21.

[24] L. Paletta and A. Pinz, “Active object recognition by view integration and reinforcement learning,” pp. 71–86, 2000.